Microbial genome evolution - the pan-genome

Pablo Vinuesa

Centro de Ciencias Genómicas – UNAM

vinuesa @ ccg_unam_mx

http://www.ccg.unam.mx/~vinuesa/

@pvinmex

Introducción a la Filoinformática:

Pan-genómica y Filogenómica microbiana –

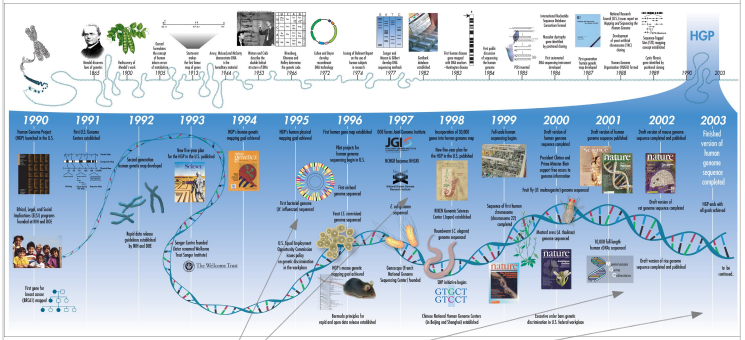
NNB & CCG-UNAM, 22-26 enero 2024

https://github.com/vinuesa/TIB-filoinfo

Genome evolution and the pan-genome

Contents

Genome sequencing and the pan-genome



1995, *H. influenzae* complete genome
1997, *E. coli* K12 (commensal)
2001, *E. coli* O157:H7 (EHEC)
2002, *E. coli* CFT073 (UPEC)

Sanger sequencing (dideoxy-chain termination) of reference genomes, mostly single individuals from different model species

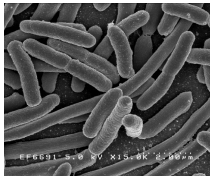
Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

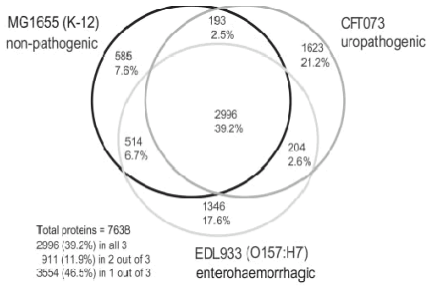
Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*

R. A. Welch*, V. Burland†‡, G. Plunkett III†, P. Redford*, P. Roesch*, D. Rasko§, E. L. Buckles¶, S.-R. Liou†, A. Boutin†**, J. Hackett*,††, D. Stroud†, G. F. Mayhew†, D. J. Rose†, S. Zhou†‡‡, D. C. Schwartz‡‡, N. T. Perna§§, H. L. T. Mobley§, M. S. Donnenberg¶, and F. R. Blattner†

17020–17024 | PNAS | December 24, 2002 | vol. 99 | no. 26

Edited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved October 24, 2002 (received for review August 30, 2002)





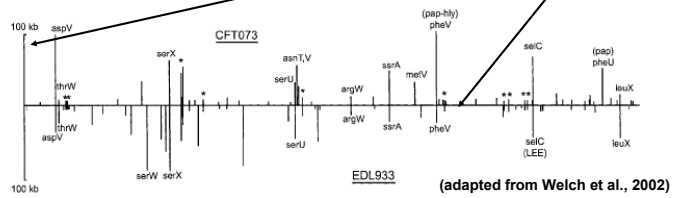
Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

"accessory/flexible genome (shell + cloud genes)":

- patchy distribution among strains
- ecological specialization
- Plasmids, phages, GIs, ICE, integrons, ISs …
- highly transferable (HGT)

"core genome":

- orthologous gene families
- species/population trees
- population genetics
- Genome annotation



(adapted from Welch et al., 2002)

Fig. 3. Locations and sizes of CFT073 and EDL933 islands. Island size, vertical axis; position in colinear backbone, horizontal axis. All islands >4 kb are shown. Islands located at tRNAs are indicated by tRNA labels. One tmRNA (*ssrA*) is also an insertion target. *, CFT073 and EDL933 islands in the same backbone location but not near tRNAs.

Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

PNAS

Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada Margarit y Ros[c], Jeremy D. Peterson[a], Christopher R. Hauser[a], Jaideep P. Sundaram[a], William C. Nelson[a], Ramana Madupu[a], Lauren M. Brinkac[a], Robert J. Dodson[a], Mary J. Rosovitz[a], Steven A. Sullivan[a], Sean C. Daugherty[a], Daniel H. Haft[a], Jeremy Selengut[a], Michelle L. Gwinn[a], Liwei Zhou[a], Nikhat Zafar[a], Hoda Khouri[a], Diana Radune[a], George Dimitrov[a], Kisha Watkins[a], Kevin J. B. O'Connor[h], Shannon Smith[i], Teresa R. Utterback[i], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,j], Dennis L. Kasper[e,j], John L. Telford[c], Michael R. Wessels[d,e], Rino Rappuoli[c,k,l], and Claire M. Fraser[a,b,k,m]

PNAS | September 27, **2005** | vol. 102 | no. 39 | 13951

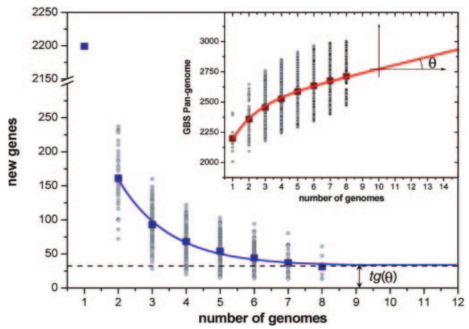
… Unfortunately, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species and also limits genome-wide screens for vaccine candidates or for antimicrobial targets.

We have generated the genomic sequence of six strains representing the five major disease-causing serotypes of *Streptococcus agalactiae*, the main cause of neonatal infection in humans. Analysis of these genomes and those available in databases showed that the *S. agalactiae* species can be described by a **pan-genome** consisting of a **core genome** shared by all isolates, accounting for 80% of any single genome, plus a **dispensable genome** consisting of partially shared and strain-specific genes.

Mathematical extrapolation of the data suggests that the **gene reservoir** available for inclusion in the S. agalactiae **pan-genome is vast** and that unique genes will continue to be identified even after sequencing hundreds of genomes.

Basic concepts in bacterial comparative genomics – Mosaic genome structure and the pan-genome

The **core** and **pan-genome** for 8 *Streptococcus agalactiae* genomes



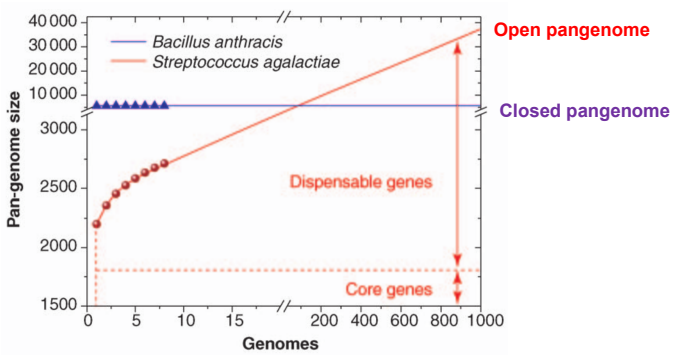
pangenome

core-genome

Species pan-genome = **core genome** shared by all isolates **+**
dispensable genome consisting of partially shared and strain-specific genes

Tettelin et al. (2005) PNAS | 102 | 39 | 13951

Basic concepts in bacterial comparative genomics – the pan-genome (Tettelin et al. 2005)



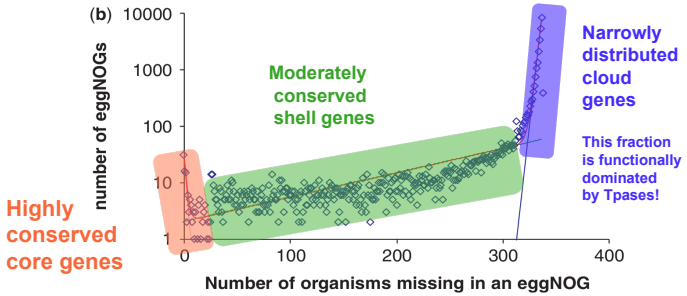
Open pangenome

Closed pangenome

Tettelin, H., et al. (2008). "Curr. Opin. Microbiol. **11(5): 472-7.**

Structure and the bacterial pan-genome

Distribuion of COGs by the number of organisms included in each cluster:

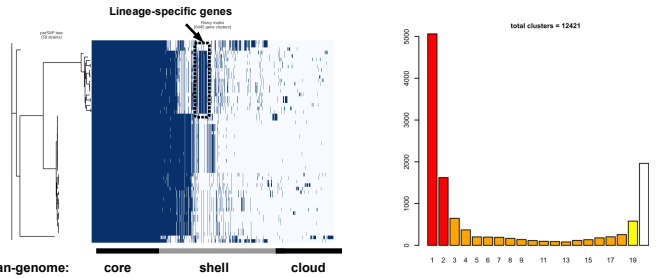


Fit of an **exponential decay function with three exponents**:

The core (~70 clusters), shell (~5700 clusters) the cloud (~ 24,000 clusters)

Koonin & Wolf, NAR 2008. Vol 36, No. 21

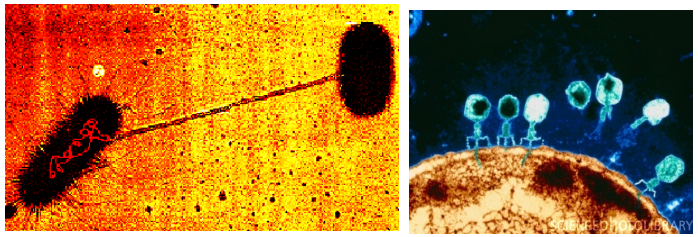
Structure of the bacterial pan-genome



Key questions in pan-genome research

- **Which are the sources and mechanisms of within-species genome content variability?**
- **How can pan-genomes be analyzed and modeled?**
- **What is the impact of population structure on pan-genome analysis and phylogenomic inference?**
- **Why do bacterial species have pan-genomes?**

Evolution of the pan-genome: prokaryotic sex (parasexuality) and HGT



- **Conjugation**
- **Transduction**
- **Transformation**

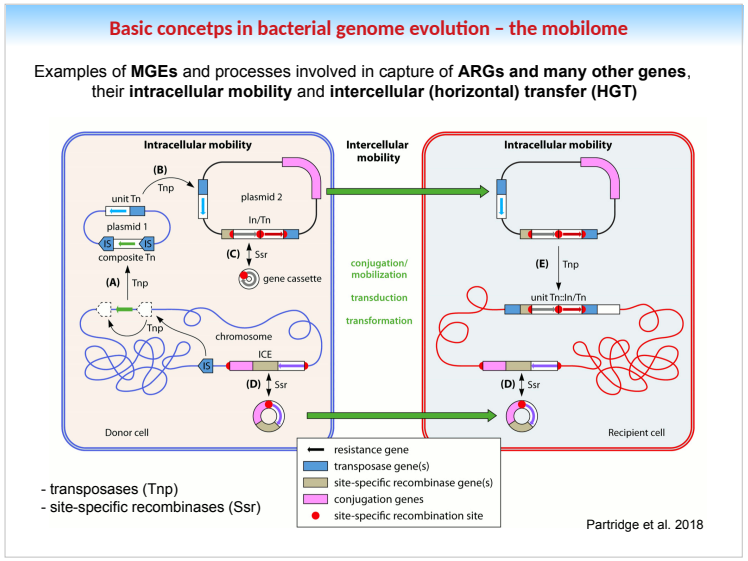
Syngamic sex vs. prokaryotic parasexuality:

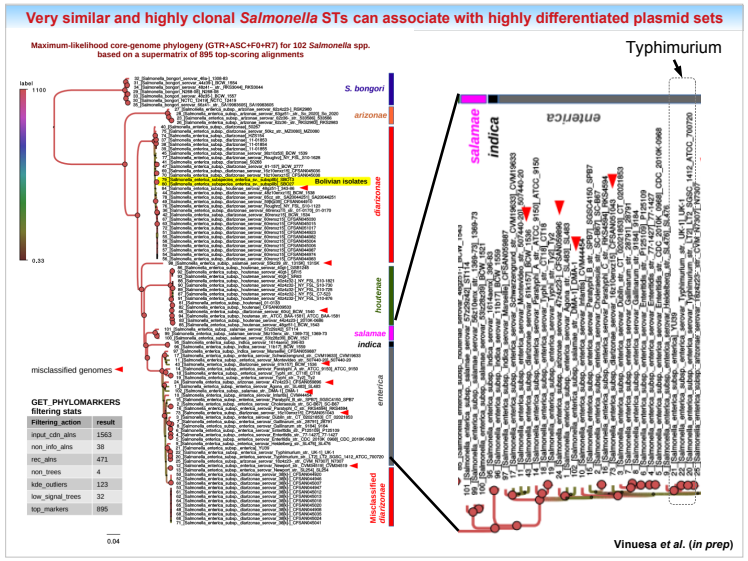
i) Bacterial sex is rare and decoupled from reproduction
ii) Can be highly promiscuous
iii) It is highly localized and mediated by different homologous and illegitimate recombination mechanisms

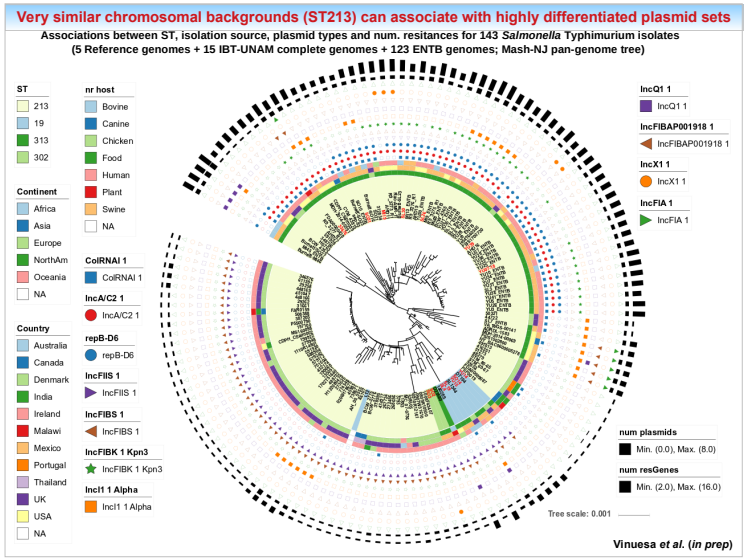
1. **homologous recomb. (*recA*)**
2. **illegitimate recomb. (transposases integrases)**

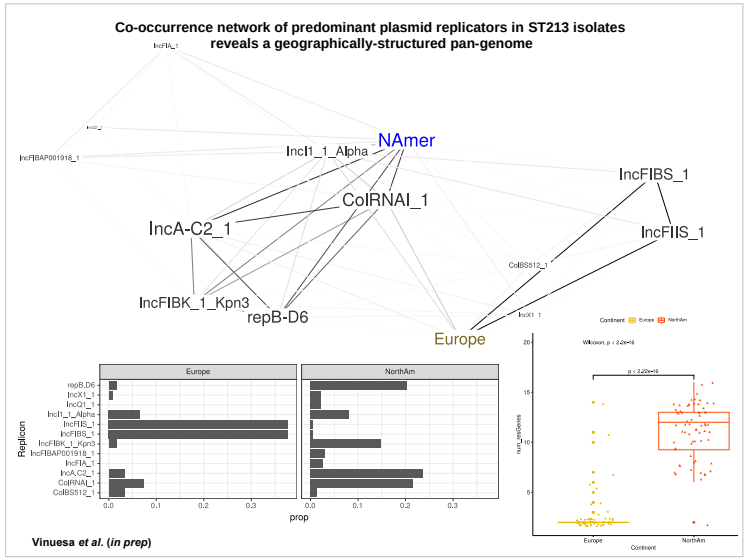
Mobile genetic elements and the accessory genome

- **Insertion sequences and compound transposons**
- **Integrons and gene cassettes**
- **Mobile and mobilizable genomic islands (GIs)**
  - **ICE: integrative and conjugative elements**
- **Prophages**
- **Plasmids**

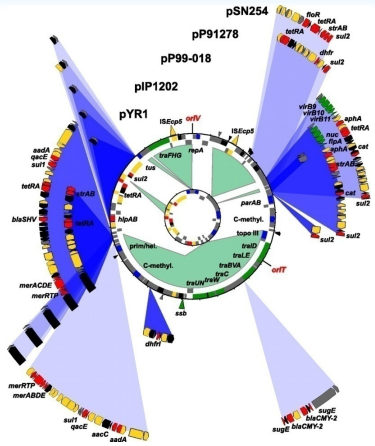


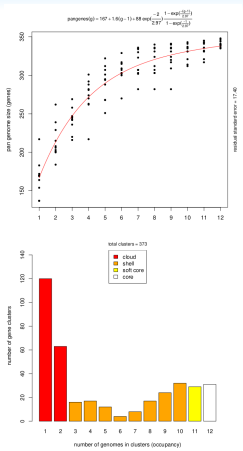






Large plasmids and other MGEs frequently have a complex mosaic structure and open pan-genome - The case of conjugative multi-resistance IncC plasmids

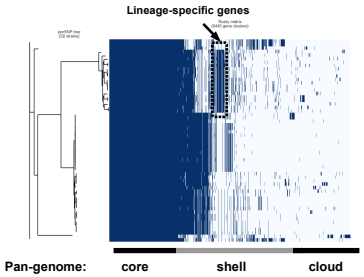


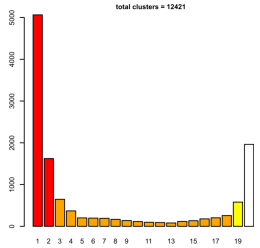


Fricke W F et al. J. Bacteriol. 2009;191:4750-4757

Vinuesa & Contreras-Moreira, 2015

Computational methods in pan-genome and phylogenomic research

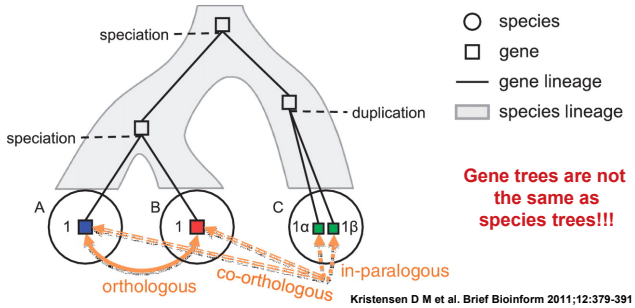




- **Inferring clusters of homologous sequences from mosaic genomes**
- **Computing and modeling pan-genomes**
- **Inferring genome phylogenies from core- and pan-genome data**

Basic concepts in bacterial genome evolution – homology relationships between gene families

Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.



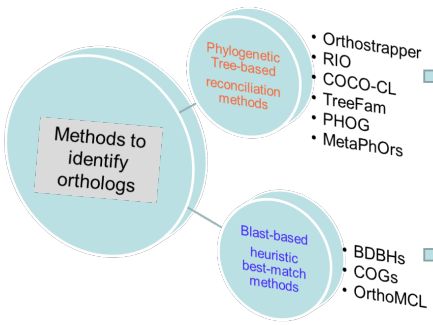
Gene trees are not the same as species trees!!!

Kristensen D M et al. Brief Bioinform 2011;12:379-391

Orthologs: essentialy instances of 'the same gene' in different species. Tend to retain function across large phylogenetic distances. Key markers for phylogenetics and genome annotation.

Paralogs: Tend to diverge over time to perform different functions via subfunctionalization or neofunctionalization routes.

Computational methods to identify homologs



- Most direct approach but known to be subjected to increasing artifacts as evol. dist. increases due to:
  - alignment problems
  - long-branch attraction
  - heterotachy
  - tree searches stuck in local sub-optimal maxima

- More adequate for bacteria, as they don't follow a clear tree-like evolutionary path (HGT).
- The only practical strategy for large datasets, therefore most common method in comparative genomics

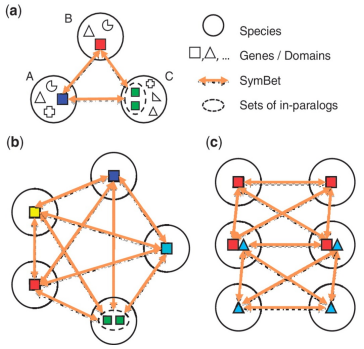
Computational methods to identify homologs

Grouping of genes in different species that are each others' BLAST BBHs into sets of orthologs and co-orthologs.

Linking pairs of BBHs from multiple genomes has a property of self-verification, as their consistency would be very unlikely due to chance, especially between phylogenetically distant lineages.

Methods for clustering of pairwise BBHs vary, but the most widely used approach involves a **single-linkage clustering procedure, where any two clusters sharing a common BBH are merged until convergence**.

Problems with: differential gene loss, domain recombination/gain/loss



Kristensen D M et al. Brief Bioinform 2011;12:379-391

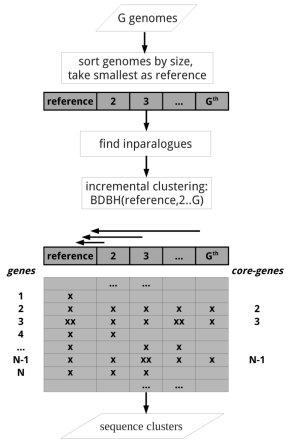
Computational methods to identify homologs: BDBHs

The bidirectional best-hit approach (BDBH)

As implemented in **GET_HOMOLOGUES**

(Contreras-Moreira & Vinuesa 2013)

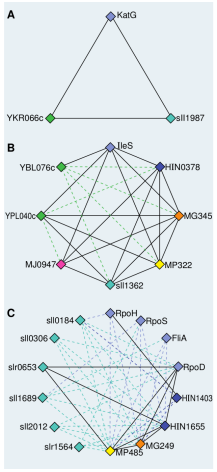
1. Find inparalogues in all genomes
2. Find reciprocal best blast hits between reference and remaining genomes
3. Cluster sequences based on filtering criteria such as:
   - % alignment coverage
   - min. % sequence identity
   - E-score cut-off
   - Pfam domain-composition
   - sinteny



Computational methods to identify homologs: COGtriangles

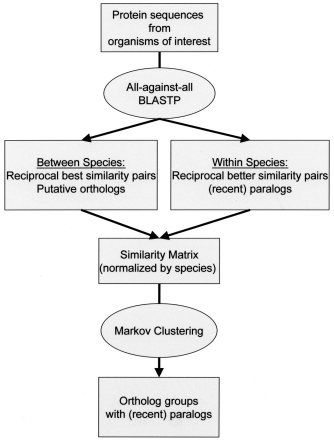
Method	Description
Clusters of orthologous groups (COGs), variants and derivatives	Identifies **three-way BBHs** between orthologs or sets of co-orthologs **in three different species**, and these groups expanded (merging triangles whenever they share a common side) until saturation, followed by manual splitting of large groups improperly joined by multidomain proteins or complex mixtures of in- and out-paralogs.

Tatusov et al. Science 1997. Vol. 278:631-637
Kristensen et al. Bioinformatics 2010. Vol. 26:1481-1487



Computational methods to identify homologs: OrthoMCL

Method	Description
OrthoMCL	Forms groups of orthologs and co-orthologs using a **Markov clustering process** involving iterative simulations of **stochastic (randomized) flow on the edges of a BBH graph**, with clusters of desired tightness identified depending on a given 'inflation' parameter determined by trial and error.



Li L et al. Genome Res. 2003;13:2178-2189

Open-source tools for phylogenomics and microbial pan-genomics
GET HOMOLOGUES & GET PHYLOMARKERS

Pablo Vinuesa
vinuesa

I'm a microbiologist interested in genomics, ecology and evolution. I develop open-source code for these topics in collaboration with @eead-csic-compbio

Edit profile

46 followers · 19 following · ☆ 10

@CCG-UNAM
Cuernavaca, Morelos, Mexico
http://www.ccg.unam.mx/~vinuesa/

Achievements

Overview · Repositories 10 · Projects · Packages

eead-csic-compbio/get_homologues — Public
GET_HOMOLOGUES: a versatile software package for pan-genome analysis
Perl ☆64 17

get_phylomarkers — Public
A pipeline to select optimal markers for microbial phylogenomics, population genetics and genomic taxonomy
Perl ☆27 7

TIB-filoinfo — Public
Talleres Internacionales de Bioinformática - Centro de Ciencias Genómicas, UNAM, Cuernavaca, México
☆6 2

intro2linux — Public
Taller de introducción al biocómputo en sistemas UNIX y GNU/Linux
Awk ☆8 3

388 contributions in the last year

Contribution activity

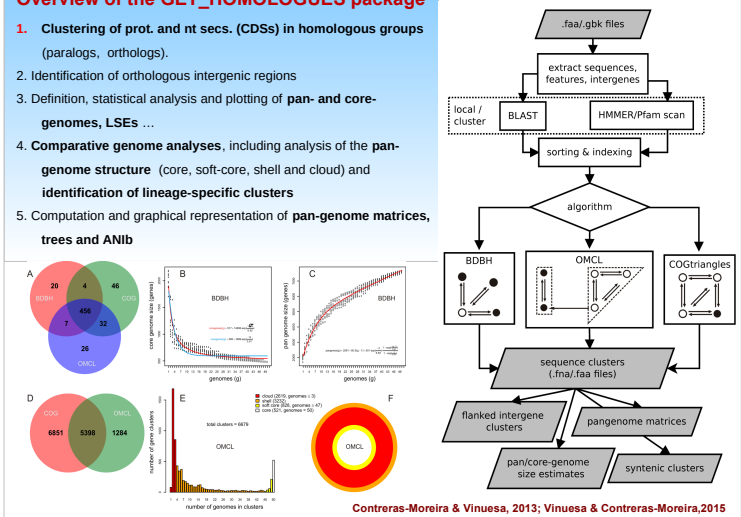
October 2021
Created 1 commit in 1 repository
vinuesa/get_phylomarkers 1 commit

September 2021
Created 87 commits in 2 repositories
vinuesa/get_phylomarkers 85 commits
eead-csic-compbio/get_homologues 2 commits

https://github.com/vinuesa

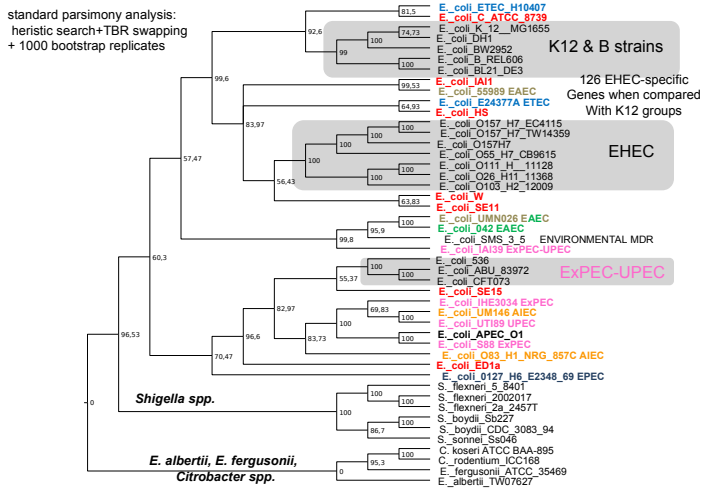
Overview of the GET_HOMOLOGUES package

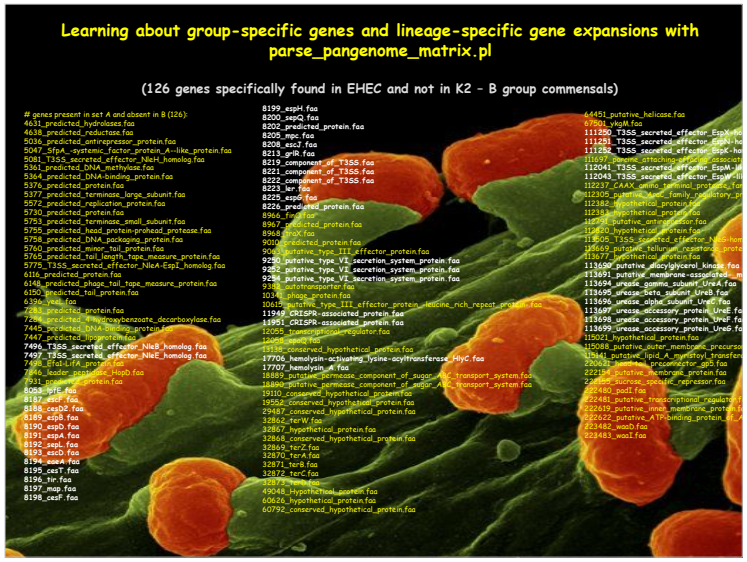
1. **Clustering of prot. and nt secs. (CDSs) in homologous groups** (paralogs, orthologs).
2. Identification of orthologous intergenic regions
3. Definition, statistical analysis and plotting of **pan- and core-genomes, LSEs** …
4. **Comparative genome analyses**, including analysis of the **pan-genome structure** (core, soft-core, shell and cloud) and **identification of lineage-specific clusters**
5. Computation and graphical representation of **pan-genome matrices, trees and ANIb**

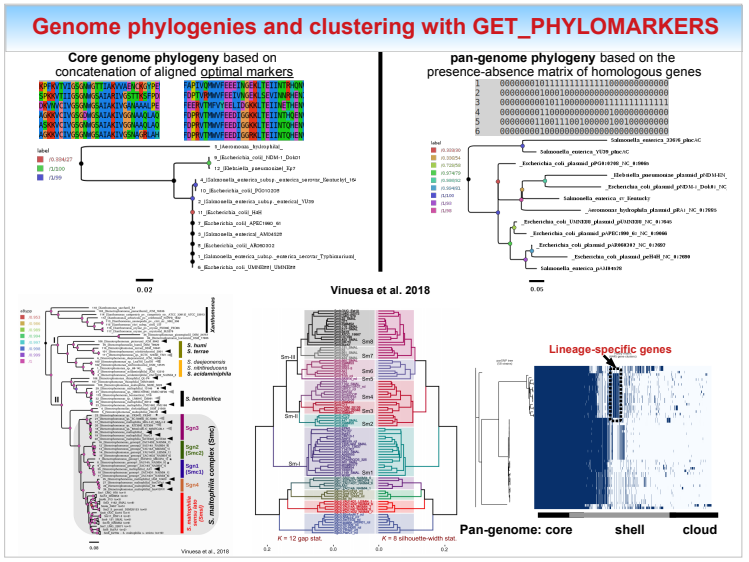


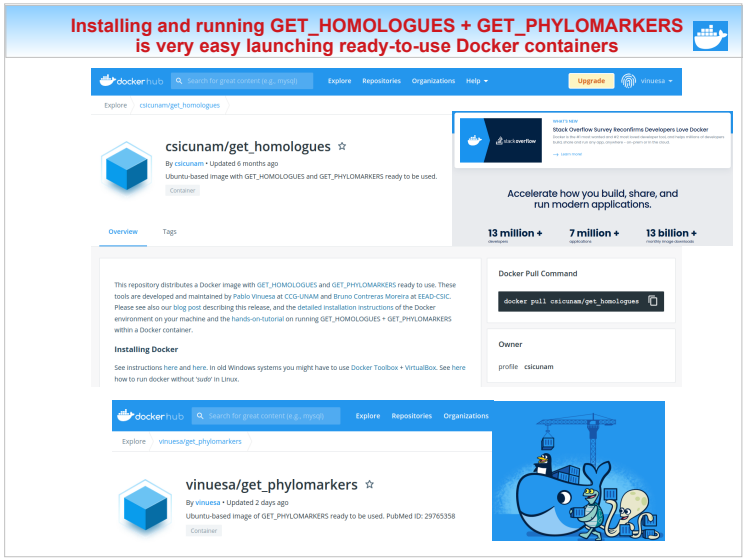
Contreras-Moreira & Vinuesa, 2013; Vinuesa & Contreras-Moreira,2015

E. coli pangenome phylogeny esimated using the consensus panmatrix of presence/absence data for 46 fully sequenced genomes of *Escherichia* spp. / *Shigella* spp. and *Citrobacter* spp.





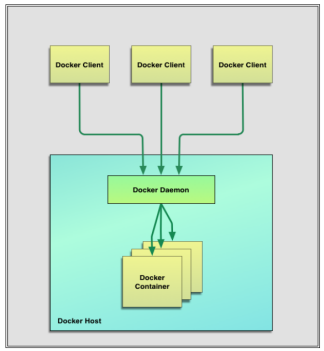


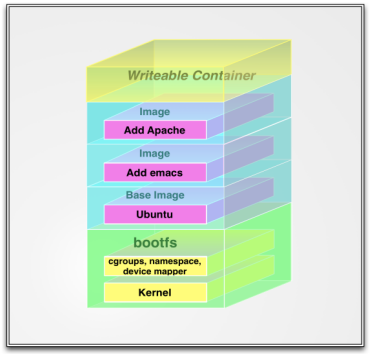


Installing and running GET_HOMOLOGUES + GET_PHYLOMARKERS is very easy launching ready-to-use Docker containers

General Docker architecture: running **Docker containers** via a **Docker client** to communicate with containers via the **Docker Daemon**

A **Docker image** is made up of **filesystems layered over each other**. At the base is a boot filesystem, **bootfs** , which resembles the typical Linux/Unix boot filesystem. Docker uses a **union mount** to add more read-only filesystems onto the root filesystem, which appear to be one filesystem.





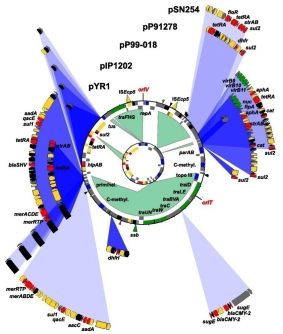
Basic Docker commands and use of containers

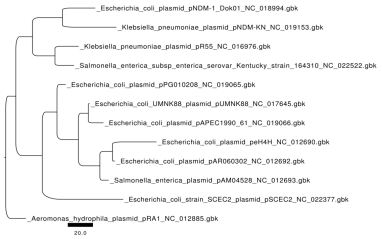
```
## To avoid permission errors (and the use of sudo), add your user to the docker group
# https://docs.docker.com/install/linux/linux-postinstall/
sudo groupadd docker
sudo usermod -aG docker $USER

# 1. Get general docker info and print help
$ docker info
$ docker --help
$ docker run --help

# 2. List available docker images on your system
$ docker image ls

# 3. List running containers
$ docker container ls

# 4. Stop a container
$ docker container stop CONTAINER-ID

# 5. Pull a Docker image from the registry
$ docker pull csicunam/get_homologues:latest

# 6. Launch an image, using a mount-bind of a user directory on the Docker container
$ docker run --rm -it -v $HOME/get_homPhy:/home/you/get_homPhy \
   csicunam/get_homologues:latest /bin/bash

# The last command uses options --rm to remove the container after exiting and sets an
-i interactive session calling a pseudo tty (-t), mounting a host directory on the
container (-v ...), accessible for wr from both, and launching a bash shell (/bin/bash)
```

Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: a case study of pIncA/C plasmids

Pablo Vinuesa[1] and Bruno Contreras-Moreira[1,2,3]
1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico.
2 Fundación ARAID, Zaragoza, Spain.
3 Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (EEAD-CSIC). Avda. Montañana, 1005, 50059 Zaragoza, Spain.

In: "Bacterial Pan-genomics". Methods in Molecular Biology. Marco Galardini, Alessio Mengoni and Marco Fondi Eds. Humana Press, Springer, 2014. *In press*





Fricke W F et al. J. Bacteriol. 2009;191:4750-4757

The GET_HOMOLOGUES + GET_PHYLOMARKERS tutorials:

- `https://github.com/vinuesa/get_phylomarkers/`

Analyses to be performed in an upcoming practical session:

A pangenomic analysis of pIncA/C plasmids using GET_HOMOLOGUES

- Defining a robust core- and pan-genome of pIncA/C plasmids
- Exploring the gene space of pIncA/C plasmids: core, shell, cloud
- Pan-genome trees vs. core genome trees (supermatrices)
- Identifying lineage-specific genes in NDM-1 producing plasmids